GENERALIZED DIGITAL CONTENT MANAGEMENT

AN OVERVIEW

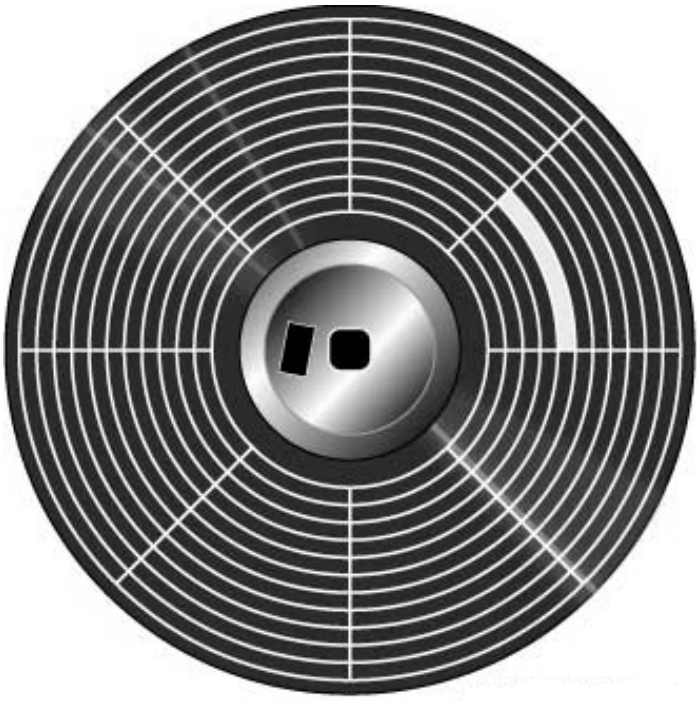
GENERALIZED DIGITAL CONTENT MANAGEMENT

AN OVERVIEW

R. GURUPRASAD
NAL, BANGALORE

ANUJ PRATEEK
BITS, PILANI

Reviewed by:
Dr Vidyadhar Mudkavi
NAL, BANGALORE



OCTOBER 2006

NATIONAL AEROSPACE LABORATORIES
BANGALORE

ACKNOWLEDGEMENT

We would like to extend our deep gratitude to Dr. A. R. Upadhya, Director, NAL and Dr. Ranjan Moodithaya, Head KTMD, NAL for granting us permission and resources for bringing out this document.

We wish to thank Dr. Khaiser Nikam, DOS, University Of Mysore, for providing very useful suggestions and thought provoking ideas to work in this exciting domain of digital content management.

We are also extremely grateful to Dr. Vidyadhar Mudkavi, Scientist, CTFD and Dr. R. M. Jha, Scientist, ALD for their kind permission and very useful suggestions to work in this very interesting area.

Finally, our thanks to Mr. Hemanth Kumar, KTMD, NAL for his in time help with the resources and editorial support.

CONTENT

ABSTRACT

This overview concentrates upon the Generalized Digital Content Management and explains various important aspects like storage media, cost elements, migration, emulation, and longevity etc-. The overview also explains various options with their pros and cons, for designing a GDCM and in the end discusses about various available organization models. The overview treats migration and emulation in the light of risk involved and focuses on various deficiencies that need to be taken care of in future. The overview will provide the reader with a deep understanding of the current trends and concepts of Digital Content management and at the same time will provide the insight into what needs to be done in the future.

KEYWORD: Cost-element, DCM, Emulation, GDCM, Magnetic Discs, Migration, Optical Discs, Risk Involvement, Storage media, etc-

INTRODUCTION

According to Peter Lyman and HAL R Varian, the world produces between one and two Exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. Magnetic storage is by far the largest medium for storing information and is the most rapidly growing with shipped hard drive capacity doubling every year. Magnetic storage is rapidly becoming the universal medium for information storage. The most striking fact to be noted according to them in their paper is the "dominance of digital" content. Not only is digital information production the largest in total, it is also the most rapidly growing.. Digital information is inexpensive to copy and distribute, is searchable, and is malleable.

In a scenario like this where the world has witnessed such exponential growth of information it is absolutely essential to develop robust content management systems through which one would look for, search and retrieve vital organizational data over long periods of time. These content management systems would invariably be huge digital repositories consisting of variety of data ranging from text, images, graphics, video, sound, documents, records and so on. Internally, the content management system would have a tightly knit algorithm for authoring, publishing, integrating, assembling, configuring, linking, delivering, sharing, searching, categorizing, transforming and archiving of information which could be accessed in the most convenient and user-friendly GUI interface as possible to the serious users of information. One could say that content management systems are synonymous with knowledge management systems. With E-Commerce gaining popularity over the Web, there is a greater demand for good content management systems. Good content management systems would address technical issues like duplication, compression, archival media, information preservation accession methods, storage formats and finally quality of service for information dissemination. Security of data is another important issue especially management of content rights and data corruption issues. A good CMS must have a well-defined “MetaData” approach, which would ensure information preservation for many years to come. Longevity in Preservation of data through MetaData tags and a very flexible hardware/software independent approach keeping pace with the technological developments are going to be crucial factors in the survival of content management systems. Obsoleteness of technology is one of the biggest enemies in terms of electronic information access. From time immemorial, the world has witnessed enormous loss of digital information, which could never be recreated due to improper archival methods. Proper ‘Migration’ and ‘Emulation’ approaches need to be incorporated in good CMS. It is also imperative to look into ‘stable’ and ‘robust’ formats standardization while designing CMS, which would last its time. Another important element which would play a huge role in designing these systems are cost elements and it is clear that the costs of preservation of digital materials will be different from other materials and will require resource commitments of a different nature on an ongoing basis. Preservation issues also need to be looked into while designing these systems as the time between an objects creation and its preservation is shrinking rapidly. It would be very apt to say here that the “costs” of preservation begin right at the time of the creation of the resource. Selection of suitable material for digital preservation and also the importance that an organization gives towards maintaining these CMS are vital factors for the survival of CMS. Some of the other cost-effective models that could be thought off are the

collaborative approaches like Library "Consortia's" which greatly help in avoiding duplication of resources and also allows sharing of resources.

A lot of work has gone in this direction for the longevity of content management systems. Some of the issues that have been addressed to preserve digital content involve methodologies like refreshing, migration, emulation, flat storage, technology perseverance and TOM. However, whatever measure we adopt we can never be assured that digital content would be preserved over an infinite period of time. Again some good practices could be followed like adoption of standards, develop strict digital preservation guidelines for everyone to follow, document the content with good "metadata" approach, have unique identifiers attached to the data, build partnerships and think of centralizing the content database. Establish adequate infrastructures even if it costs a bit high. Never be "Penny Wise and Pound Foolish" especially while building good CMS. In this overview, we have also included risk management of migration and emulation is also addressed. Towards the end of this document, we have also highlighted various organizational models, which are prevalent today. Hope you would enjoy reading this overview.

SECTION - I

Generalized Content Management System
Content Amount and Type
Issues and Needs
Cost Element

1. GCMS

In the modern information era, the volume of information has increased exponentially and with it has grown the information overload. It is required to present this information to the user in a well-defined and accessible manner, restricting the use at the same time governed by certain policies. This has given rise to a new objective called as content management system. In simpleton language, one can understand content management system as a system that manages content. In the context content is the information, which may include anything, ranging from text, images, graphics, video, sound, documents, records etc – Here we concentrate on electronic content mainly since it is suffering with the most serious onslaught of information overhead. Moving a step further we define content management system (CMS) as a tool that enables a variety of (centralized) technical and (de-centralized) non technical staff to create, edit, manage and finally publish (in a number of formats) a variety of content (such as text, graphics, video, documents etc), whilst being constrained by a centralized set of rules, process and workflows that ensure coherent, validated electronic content that is easily accessible and understood by the user of the particular content.

Broadly speaking CMS primarily concentrates at authoring, acquiring, publishing, dynamic page generation, integrating, assembling, versioning, configuring, linking, delivering, caching, analyzing, sharing, searching, categorizing, transforming, re-using, syndicating, archiving, etc of the content but at the same time, CMS has to meet the demand of repositories for the content for future use.

Various analysts have divided CMS into various groups and they are in a more or less similar way applicable to GCMS too. Gartner divides CMS into,

- Enterprise internal content: This includes corporate portal, document management, media asset management, retrieval, software configuration management, and product data management products.

- Web site content: This includes web publishing and document management products.

- Shared content: This includes business partner and supply chain collaborative content that is not transactional.

In the same manner, GIGA categorizes CMS into following four parts,

- Software configuration management (e.g. MKS)

- Document management (e.g. Documentum)

- Web publishing (e.g. EBT), and

- E-commerce servers (e.g. Blue Martini).

Stepping ahead of others, CAP Ventures' view about CMS that CMS is an "umbrella" that includes numerous technologies including, but not limited to,

- "Web Publishing: Tools used to create, manage, and deliver content to the web, including digital asset management systems.

- Collaboration: Tools used to assist content-driven creation and communication processes, including document management systems.

- Portals: Tools that provide a window to a wide range of information, from company documents and data to external resources and publications.

- Content Enrichment: Tools that help refine the selection of content for delivery or viewing, including taxonomy, categorization, personalization engines, and analytical tools, etc.

- Content Distribution: Tools involved in managing content outside the direct control of the dynamic content technology system, including syndication servers and digital rights management systems."

This view of CAP Ventures' marks the growth of idea of GCMS, which is one of the most current topics of discussion in CMS.

It is predicts that CMS market share will cross $50 billion by end of this decade and it clearly shows importance of CMS and this led to the evolution of Generalized CMS (GCMS) from web, document, knowledge, digital asset etc management systems. Another reason for GCMS management evolution is e-commerce, which has increased drastically in recent years. The market information comes from exhaustive study by various perspectives as vendor revenue, analyst view, user view etc- and it clearly shows that this area needs more attention and at the same time seeing the market dynamicity and progress, its very arguable that in vary near future GCMS will be the next line of front.

Defining GCMS on the line of CMS the only difference that comes into picture is the way we treat content. GCMS doesn't distinguishes between various types of content, rather in terms of today's technology, either sees it as all binary content of goes above a general level of abstraction to call it objects. The result of it is in GCMS we do not need to distinguish between various CMS types rather adding automation we can deal with any kind of data in a common manner.

2. CONTENT AMOUNT & TYPE

GCMS's target is the content so it is necessary to analyze the amount of content GCMS expects to handle. Speaking in general, all the content that is available to any human being comes under its scope, this leads to a capacious amount of content in front of the GCMS, and the very first consequence of it is the storage problem. The world produces between one and two exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. An exabyte is a billion gigabytes, or 10^{18} bytes. Printed documents of all kinds comprise only .003% of the total. Magnetic storage is by far the largest medium for storing information and is the most rapidly growing with shipped hard drive capacity doubling every year. The table given below briefs the content in front of GCMS,

Storage Medium	Type of Content	Terabytes/Year, Upper Estimate	Terabytes/Year, Lower Estimate	Growth Rate, %
Paper	Books	8	1	2
	Newspapers	25	2	-2
	Periodicals	12	1	2
	Office documents	195	19	2
	Subtotal:	**240**	**23**	**2**
Film	Photographs	410,000	41,000	5
	Cinema	16	16	3
	X-Rays	17,200	17,200	2
	Subtotal:	**427,216**	**58,216**	**4**
Optical	Music CDs	58	6	3
	Data CDs	3	3	2
	DVDs	22	22	100
	Subtotal:	**83**	**31**	**70**
Magnetic	Camcorder Tape	300,000	300,000	5
	PC Disk Drives	766,000	7,660	100
	Departmental Servers	460,000	161,000	100
	Enterprise Servers	167,000	109,000	100
	Subtotal:	**1,693,000**	**635,660**	**55**
TOTAL:		**2,120,539**	**693,930**	**50**

Table 1. Content type, storage medium, estimate of amount and growth.

The striking fact that emerges from the table is the dominance of the digital content. Digital information production is the largest in total and it is also the most rapidly growing. While unique content on print and film is hardly growing at all, optical and digital magnetic storage shipments are doubling each year. Most textual information is born digital, and within a few years, this will be true for images as well. These facts probably owe to the inexpensiveness of the digital data in terms of

creation, copying, and distribution. Moreover, the digital data is easily searchable and is malleable. These facts forces GCMS to concentrate on digital data and at the same time forces the designer to address the storage, retrieval and search policies more efficiently.

The amount of content that the table points towards is not very accurate and before the real GCMS is designed, needs to be refined to get the better picture about the amount. The content that GCMS handles is mostly refined based on the following filtering parameters,

- Duplication: Remove any data that exists in one or more copy or is very similar to other.

- Compression: Many data may be in compressed format and this yield to a error in the judgment about the size of the data. Even if some compression is used, it needs to be standardized.

- Archival Media: A lot amount of the data is stored on devices like tape drives in form of backup and it is required to judge if all the amount of historical archives be included into the scope of GCMS.

Next thing that becomes of demure concern is accessibility. To answer this issue we need to see how the data flow is happening. The major information flow is found to be happening from telephone, web, and emails, TV, radio etc, and this force another concern of security to be addressed. Apart from that, another important thing to consider is about the data consumption. The table given below summarizes the data consumption and this table forces GCMS to focus on security.

ITEMS	1992 Hours	2000 Hours	2000 MB	%Change
TV	1510	1571	3,142,000	4
Radio	1150	1056	57,800	-8
Recorded Music	233	269	13,450	15
Newspaper	172	154	11	-10
Books & Magazines	185	176	13	-10
Home video	42	55	110,000	30
Video games	19	43	21,500	126
Internet	2	43	9	2,050
Total:	**3,324**	**3,380**	**3,344,783**	**1.7**

Table 2. Yearly data consumption.

To have a look on how the information is growing, presented below is the yearly production of the information,

Item	Titles	Terabytes
Books	968,735	8
Newspapers	22,643	25
Journals	40,000	2
Magazines	80,000	10
Newsletters	40,000	.2
Office Documents	7,500,000,000	195
Cinema	4,000	16
Music CDs	90,000	6
Data CDs	1,000	3
DVD-video	5,000	22
Total:		**285**

Table 3. Yearly information production.

It is evident from the table that the digital content is produced in a very large volume and the rate of production is very high and will pose a big challenge in front of GCMS.

To summarize, we can easily argue that the GCMS is to concentrate mainly on digital data, which owes to the enormous fraction it has and if we need to distinguish the data then multimedia data is going to be the main concern in terms of accessibility and storage.

3. ISSUES & NEED

The main issues that GCMS faces are surely due to massive amount of data and the major fraction of multimedia data in the content. The solution can be found more comfortably if we break the problem into much smaller parts and list given below shows some of the thought sub-problems,

- Data documentation: This eases the search of desired information.
- Format of storage: To answer future availability of the data.
- Storage medium: To answer technological aspects at a cheaper price.

Designing such a system involves certain technical issues also along with the conceptual & theoretical problems and a list of few of them is given below,

- Storage, organization, and management of the system through software available and there selection or development is not easy.
- Available physical bandwidth in the delivery path to the users falls short.
- Quality-of-service (QoS) management like real-time delivery and adaptability to the environment are necessary.
- Information management (indexing and retrieval) in an automated manner needs advanced algorithms and concepts.
- User satisfaction.
- Security, especially management of content rights and data corruption.

If we separate the management part from GCMS, what we are left is the repository and we list few requirements that the repository must satisfy,

- Storage of the content,
    - Metadata and various attributes need to be present.
    - Any format should be acceptable to the repository.
    - Provide different storage areas with different access methods for different kinds of content.
- Flexible access to the content,
    - By proper documentation of in a generalized manner.

  - By proper indexing schemes.
  - By providing search interfaces and methods.
  - By offering filtering and scaling methods for retrieval according to available bandwidth or location.

- Others,
  - Decentralized implementation with transfer and replication mechanisms.
  - Remote access facilities.
  - Integration of rights management.

To provide the desired interfaces certain modules require to be in place and an example list is given below,

- To manage a large amount of content interactively.
- To provide a centralized access to all stored content.
- To preserve physical quality of the content.
- To provide functions for content based searching, content browsing and retrieval.

Designing such repository leaves us to ponder on various design issues to be considered. A list of such issues is given below,

- Server subsystems that store, manage, and retrieve the multimedia data streams upon user request in an efficient manner are hard to establish and select.
- Network subsystems that transport, delivers, adapt and transform the data streams isochronously, to the clients without loss and demur.
- Client subsystems that receive the data streams and manage the presentation of data in an understandable manner.
- Application programs that deal with relationships among data frames and media segments, and manage user navigation and retrieval of this data through a user-friendly interface.

The issues of implementation and design owe to the hardware limitations and lack of very efficient technological algorithms. The issues are further graven by lack of standardization.

The development of GCMS, which would be able to work really well in all aspects, requires a lot of work force and money but we have sufficient reasons to spend our resources on it. A few of those reasons are listed below,

- An enormous amount of digital information is already lost and can never be recreated and the only reason behind is improper archival. Due to improper archival format now, a day, we do not have access to many of the information, which we have and are altogether left inaccessible. The improper selection of medium of storage also led to heavy loss. If only proper data format and longevity issues were taken into consideration, the data would have been present with us.

- There will be a demographic bulge of electronic materials coming into libraries and archives as the new information will come into picture. To make our future task better we need to organize and group our data today with help of various schemes of metadata and attribute retrieval.

- Information technology is growing at a very high speed and this leads to obsoleteness of technology very fast, so if the proper migration and preservation schemes are not developed our information will become inaccessible tomorrow.

- Since both print and digital content are increasing very high the discrimination between these two types of content needs to be removed and addressed as it.

- Due to heavy information hiding due to weakness in IP, laws will make the content and information in accessible and the GCMS is the only option that can centralize the IP law to help the contents authenticity and security by providing a common format.

- Due to lack of standards, it is becoming increasingly difficult to automate data retrieval, organization and is leading to data loss.

Keeping all the things in mind the question that needs to be answered is that what needs to be done now? Briefly, we can list the following points that are given below,

- Creation and collection of knowledge.

- Selection of content to be saved.

- Ensure vital electronic documents are preserved now.

- Document formats standardization.

- Being legal: Rights management and access control.
- Promoting and implementing preservation techniques.
- Digital preservation for public good i.e. to make major content base open source.

Digital collections facilitate access, but do not facilitate preservation. Digital places greater emphasis on the here-and-now rather than the long-term, just-in-time information rather than just in case. With so many positive attributes attached with the digital information it becomes very important to create a system to manage it and the answer is obviously GCMS.

To speak further we boil down our main concerns from our experiences about CMS that are the prime suspects that will affect GCMS also. A list of such suspects is shown below,

- Longevity.
- Format of storage.
- Medium of storage.

Proceeding in the way we hope that we would be able to analyze and review all the problems that GCMS may encounter and at the same time may be able to propose some solutions. Furthermore, we would consider digital preservation as our key focus as that is the most important aspect of GCMS.

4. COST ELEMENT

It is anticipated seeing the volume and discreetness of the information we have that the digital preservation after implementing GCMS will take many resources in term of work force or money. Although it may too early to make meaningful comparisons of the costs of digital vs. traditional preservation, one thing is certain: the costs of preservation of digital materials will be different from for other materials and will require resource commitments of a different nature on an ongoing basis.

- The important points to be noted here is that the costs of preserving digital materials depend on various factors like,
- The list of elements that get into the collection manager's workflow,
- Costs for preservation cannot be separated from costs of access,
- An institutions investment on technical infrastructure (cost is shared for both preservation and access),
- Costs for providing resource discovery and information retrieval (delivery of material) from the archive also depend on the extent to which the archive is integrated into the collection management functions.

4.1. TIME FRAME IN DIGITAL PRESERVATION

The cost of preservation of digital material is an ever-ongoing commitment irrespective of whether the preservation is digital or relates to traditional materials. The time between an objects creation and its preservation is shrinking rapidly. Preservation will need to be addressed increasingly at the time of acquisition or even creation of the digital resource. For the newer digital materials it is still not clear as to what would be the long-term commitment. It depends on various factors and few of them are listed below,

- The archiving model that one wishes to adopt,
- The technical strategy chosen for both preservation and access, and finally
- What would be the optimum migration strategy that needs to be adopted?

4.2. THE LIFE CYCLE OF A DIGITAL RESOURCE

Digital materials are created only to require some sort of ongoing "re-creation" (Migration, refreshing onto new media etc.) In order to ensure access is preserved. As far as digital materials are concerned, the link between its creation and preservation is much more important because decisions about the way a digital object are created influences how (or indeed whether) it can be preserved. Decisions taken at the time of preservation can have a tremendous impact in the end as eventually this information has to be retrieved in its proper format. In all libraries wherever digital projects are

initiated, preservation strategies must be thought off and properly documented as early as possible. One could say that the "costs" of preservation begin right at the time of the creation of the resource. Creation of a digital object is the true starting point for its preservation.

4.3. THE COST BENEFIT TRADE-OFF

Digital preservation would inevitably be about trade-offs. A robust preservation policy combined with the easy retrieval mechanisms is the order of the day. Spending enormous costs to store a complex digital object to which no one requests for its access is also not desirable. The preservation strategy should be appropriate to the perceived value of the digital object. The benefits of preservation are inextricably linked with the policies for selection of the right material for archiving.

4.4. MATERIAL SELECTION FOR DIGITAL PRESERVATION

In selection of material for digital preservation, there do exist management policies that need to be kept in mind within an institution, how suitable is the object for preservation and technical factors concerning the specific digital object and its requirements for long-term access. These factors even though listed individually must be looked into as a cohesive whole.

Collection Management Policy Issues: As far as digital materials are concerned, creation/acquisition and preservation are inextricably linked and decisions about preserving materials for the long term should reflect selection policy for the collection as a whole.

- Technical Considerations: Ultimately, everything boils down 'bits and 'bytes' in computer jargon. Then, what meaningful readable information that could be extracted from these is the crux of the whole issue of preservation. With regard to digital materials, simply maintaining a byte stream does not necessarily ensure the digital material will be preserved at a level acceptable to the archive and its user. For digital materials, access can be at various levels right from the full range of the functionality or content to simply access the 'bare bones' of the intellectual content. The levels at which a digital material is archived and maintained depends upon many technical judgments of the archivist. "Metadata" (or in other words "representation information") is nothing but determining the significant properties of a digital object. Generally, for all digital materials the preservation of complex functionality may prove considerably more costly than preservation of the basic intellectual content. More complex the digital object, more intensive would be the preservation methodologies. Migration is another alternative to reduce cost. Here, one should adapt standards or system independent file formats during creation or during migration. The idea is to maintain access over long periods.

- Collaborative Approaches: Library consortia are becoming a welcome solution to cut down costs and avoid duplication of efforts across the various libraries in the world. Collaborative efforts could significantly reduce the cost for Organizations. One thing to be kept in mind here is that the cost factors may vary depending on whether the collaboration is occurring regionally, nationally, or internationally.

Generally selection decisions are based on existing policy documents or existing organizational policies, wherever there are no pre-defined practices, then it is taken object by object or on a collection-by-collection basis. Obviously, more time is taken if there are no pre-defined or existing policies in place. If there are well-defined collaborative agreements like consortia's, then the whole preservation effort could be less time consuming and cost effective.

One has to ensure that the digital object is adequately prepared for archiving as well as the resources for agreeing on a specific preservation strategy for continuing access (could be migration or emulation). A detailed consideration of the digital object is essential to determine its significant properties.

4.5. STORING OF FILES

This would include maintenance of hardware, software, and transfer of files from one generation of storage media, periodic inspection of stored files and of the storage media itself. Costs for taking backup copies need also be considered.

4.6. ADMINISTERING THE ARCHIVE

Developments in the technology, the prevailing law of the land would make a significant impact in preservation of the archive and its periodic updation. Costs may also include changing the archive system in accordance with the changes in the archiving policy.

Most importantly, digital archiving should include staff costs (salaries, training, re-training and upgradation of skills), insurance, building overheads, certification etc.

REFERENCES

In this section, the following technical literature and web sites were referred to:

I. *"What Is Content management?"*, The Gilbane Report, Vol. 8, No. 8, Oct. 2000.

II. Besser, Howard. *"The Changing Museum"* in Ching-chich Chen, ed., Information: The Transformation of Society, pp. 14-19. Proceedings of the 50th Annual Meeting of the American Society for Information Science, Medford, NJ: Learned Information, Inc.

III. P. Lyman & H. R. Varian, *"How Much Information?"*, JEP, Vol. 2, Issue 6, August 2001.

IV. Alision Bullock, "Preservation of Digital Information: Issues and Current Status", ISSN 1201-4338, Information Technology Services, National Library of Canada, August 1999.

V. A. Dan, S. I. Feldman & D. N. Serpanos, *"Evolution And Challenges In Multimedia."*, IBM Journal Of R&D, Vol. 2, Nov. 1998.

SECTION - II

Digital Preservation
Storage Medium
Risk with Migration and Emulation
Formats of Storage

5. DIGITAL PRESERVATION

"Digital preservation" or "digital archiving" means taking steps to ensure the longevity of electronic documents. It applies to documents that are either "born digital" and stored electronically or to the products of analog-to-digital conversion, if long-term access is intended and this is one of the major concerns of GCMS as was with CMS. Any data or content that we consider in GCMS is taken in terms of binary data and a bit stream can be stored in many different ways on different media. Retrieving a bit stream from its physical representation on some medium requires a hardware devices, as well as special “controller” circuitry that can retrieve the information stored on the medium in an understandable format. A special program called as device driver is also required to make this device accessible by a given computer system or digital device.

A bit stream has implicit structure that cannot be represented explicitly in the bit stream itself. A bit stream represents a sequence of data generally as fixed-length chunks of information (called “bytes”), each of which represents a code for a single data unit. In current schemes, bytes are typically 7 or 8 bits long. However, a bit stream cannot include enough information to describe how it should be interpreted. In order to extract fixed-length bytes from a bit stream and in principle we encode a integer called as the key in the beginning of the bit stream, representing the length of each byte. However, this key integer must itself be represented by a byte of some length else the user cannot interpret the key and hence we need another key to explain how to interpret the first key. This leads us to a recursive problem that is solved by help of a bootstrap. In order to provide such a bootstrap, we must annotate our digital storage medium with easily readable information that explains how to read it. Yet this leads to a problem similar to that of encoding a key to specify the length of each byte in a bit stream. To interpret each byte, we need to know what coding scheme it uses; but if we attempt to identify the coding scheme by encoding a “code-identifier” in the bit stream itself, we need another code-identifier to tell us how to read the first code-identifier. Again, we must bootstrap this process by providing easily readable annotations.

Digital documents have the discouraging characteristic of being software-dependent. They can by using appropriate software’s only and it is necessary to run the specific software that created a document. This point to the storage of the software along with the digital document, else afterwards in future our documents may become unreadable.

Preserving digital documents is analogous to preserving ancient written texts. Just as with digital documents, it is sometimes necessary to refresh an ancient text by transcribing it, since the medium on which it is recorded has a limited lifetime. The solution to it is migration. Example of lifetime of some commonly used medium is given next,

Device	10°C	15°C	20°C	25°C	28°C
D3 magnetic tape	50 years	25 years	15 years	3 years	1 year
DLT magnetic tape	75 years	40 years	15 years	3 years	1 year

cartridge					
CD/DVD	75 years	40 years	20 years	10 years	2 years
CD-ROM	30 years	15 years	3 years	9 months	3 months

Table 4. Common storage devices and expected lifetime at various temperature.

Copying text in its original language guarantees that nothing is lost assuming that knowledge of the original language is retained along with the text. This amounts to saving the "bit stream" of the original text but sometime this task may also become very difficult if the medium is not accessible.

One solution to it seems to translate digital documents into standard forms that can be guaranteed to be readable in the future. This would circumvent the need to retain the ability to run the original software that created a document or to ensure we are following some available standard format for the document like relational database, XML, SGML etc. Files represented using some standard that will be available afterwards also be copied to new media, as necessary, and the standard would provide readability for all time.

Other issue that needs to be taken care of is hardware, if the hardware runs obsolete then the content will become inaccessible in future too and for this, we need to periodically transfer or migrate our contents from old medium to a newer one.

Well the problems that we face in the preservation of digital media can be attributed to some indirect issues, which are in other sense the deterrent in the path of development of a GCMS that solves the problems of the longevity. The problem that we face here can be summarized into various aspects in light of social, technical, and legal scenario as follows in the points given below,

- The rapidly increasing number of digital objects and proliferation of document standards and formats make it difficult to standardize the storage.
- The increasing complexity of digital objects and their increasing software dependence.
- The lack of planning to incorporate preservation needs in systems and lack of availability of off-the-shelf products supporting preservation needs.
- The lack of consideration of long-term access requirements when creating digital products.
- The absence of widely accepted standards, which will assure access over time.
- Copyright/intellectual property rights that may interfere with the ability to preserve digital objects through systematic copying.
- Unstable storage media whose life span is limited.
- A lack of technical expertise in collections managers and preservation experts.

- An emphasis on the creation and/or acquisition of digital material in an era of diminishing resources, rather than ongoing preservation and access to existing electronic holdings.

In light of the problems mentioned we can also summarize what we need to do and a list shown, lists them below,

- Fix the object as a discrete whole i.e. object storage.
- Preserve the physical presence, as the perseverance of a physical file does not guarantee its accessibility.
- Preserve content i.e. maintaining the ability to access the content at its lowest level, such as ASCII text, without the embellishments of font variations and layout features.
- Preserve the presentation
- Preserve functionality.
- Preserve authenticity and security.
- Preserve provenance i.e. assert the origin and chain of custody of an object and contributes to defining it as a whole and preserve it.
- Preserve context i.e. the dependencies.

For preserving any digital content, based on the requirements and issues that we talked about, certain methodologies are given and are listed here,

- Refreshing: It involves periodically moving a file from one physical storage medium to another to avoid the physical decay or the obsolescence of that medium. Because physical storage devices decay, and because technological changes make older storage devices inaccessible to new computers, refreshing is necessary.
- Migration: It is an approach that involves periodically moving files from one file-encoding format to another that is useable in a more modern computing environment. Migration seeks to limit the problem of files encoded in a wide variety of file formats that have existed over time by gradually bringing all former formats into a limited number of contemporary formats.
- Emulation: That is create a virtual interface between the software and the hardware and then even if the hardware is changed then the virtual interface can be built to work on that and the software working on a higher abstraction level will be able to work as if working on the older hardware. So, basically by

emulation we will be providing backward compatibility to the content. The diagram given next explains emulation,

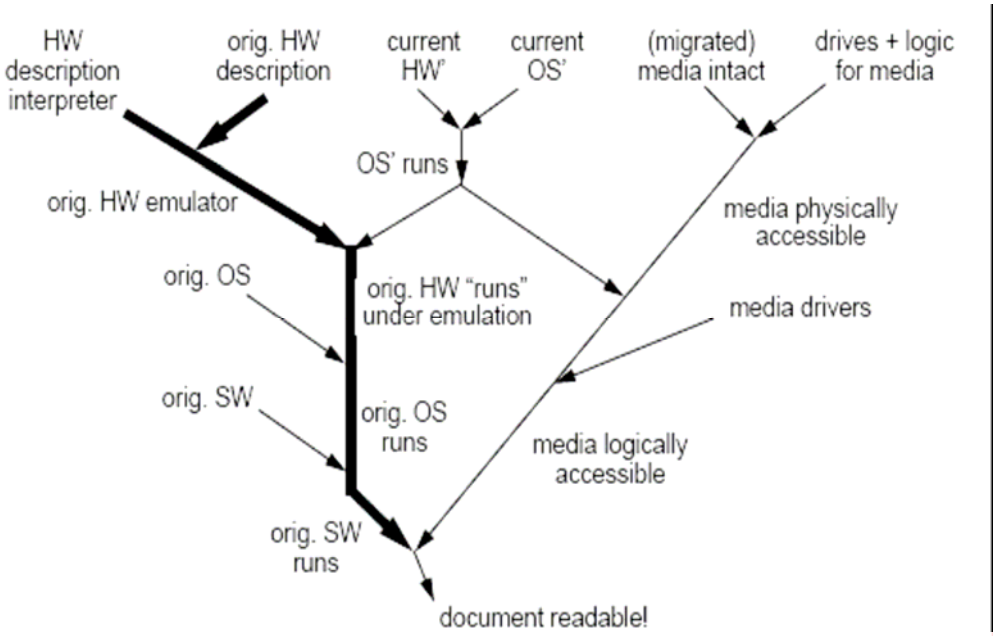


Figure 1. Emulation.

- Flat storage: Since, paper and films are known to have the longest storage we can save the documents on them. Here the problem remains is how to store multimedia, but a proposed solution is binary storage. The problem that will remain unsolved is do we have enough flat media to store the digital content?

- Preserve technology: Another method for ensuring ongoing access to digital objects would be to keep older technology available for use. Although this would preserve content and enable future generations to view digital objects in their native format with original layout and functionality, creating hardware or software "museums" is prohibitive in cost, space, and technical support requirements. At best, this method is an interim measure when migration is not possible.

- TOM: This starts out with the recognition that all digital data things are objects, that is, they have specified attributes, specified methods or operations, and specific semantics. All digital objects belong to one or another type of digital object, where “type” is defined by given values of attributes, methods, or semantics for that class of objects. Any digital object is a byte sequence and has a format, i.e., a specified encoding of that object for its type. Byte sequences can be converted from one format to another. The content and appearance of the document remains identical.

The diagram given next shows various preservation methods and the categorization,

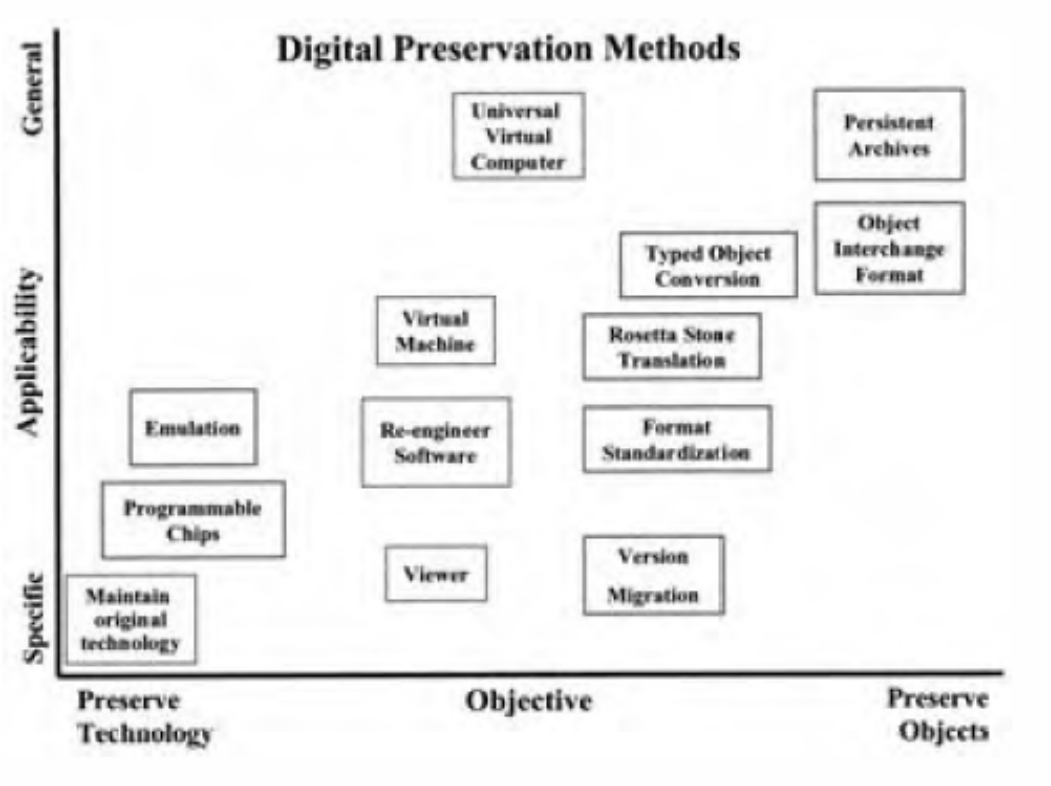


Figure 2. Digital Preservation Methods.

A lot of research is undergoing on the above lines. The National Preservation Office and the Joint Information Systems Committee (JISC) co-funded several projects on digital archiving. One outcome of this work is a tool for measuring the complexity of the preservation process and guiding selection of a preservation approach. The JISC is building on this foundation CEDARS 3 project through the Consortium of University Research Libraries. The three-year project began in March 1998. Among other objectives, CEDARS will investigate methods of preserving different sorts of digital resources and develop priced and scaleable models. Cornell University is working to create risk management tools for the management of digital information, and to develop a plan for the long-term preservation of Cornell's digital documents.

Well we can never argue with whatever measure we follow the digital content will be preserved for infinite time, but we can follow some methods that may increase the probability of long life of the digital media. The list shown below, lists a few of them,

- Adoption of standards: This means to adopt a standard that is most commonly used and keep a track of how it is changing and migrate to the newer one, as it requires.

- Develop digital content preservation guidelines, mandatory for everyone to follow.

- Document the content.
    - Metadata attachment and generation should be done.
    - Unique identifiers must be attached to the data.
- Building partnerships and centralizing the content database.
- Establish infrastructures even if it costs a bit high.

In light of whatever we have said until now we can say that we must develop evolving standards for encoding explanatory annotations to bootstrap the interpretation of digital documents that are saved in nonstandard forms and must develop techniques for saving the bit streams of software-dependent documents and their associated systems and application software. At the same time, we must ensure that the hardware environments necessary to run this software are described in sufficient detail to allow their future emulation. Specifications, annotations must also be saved as with digital documents and the contextual information should be associated with it. Finally, we must ensure the systematic and continual migration of digital documents onto new media, preserving document, and program bit streams verbatim, while translating their contextual information as necessary.

6. STORAGE MEDIUM

GCMS is a concept, its realization depends upon certain hardware, and hence it becomes one of the most important points of concern when we talk about GCMS. In hardware the most important part is the storage medium as it attributes to the longevity of the content, its access etc. Storage leads to certain storage practices relate to the plans for migrating from current hardware and software environments to newer environments, the refreshing of media, and backup and recovery. Here since digital content and storage is talked the only storage type that comes into picture is computer storage. Computer storage, computer memory, and often casually memory refer to computer components, devices, and recording media that retain data for some interval of time. Here we will limit ourselves to various aspects of computer storage.

Some important types of computer storage mediums are listed below,

- Magnetic storage: Magnetic storage uses different patterns of magnetization on a magnetically coated surface to store information. Magnetic storage is non-volatile. The information is accessed using one or more read/write heads. Since the read/write head only covers a part of the surface, magnetic storage is sequential access and must seek, cycle or both. In modern computers, the magnetic surface will take these forms,
    - Magnetic disk.
    - Floppy disk, used for off-line storage.
    - Hard disk, used for secondary storage.
    - Magnetic tape, used for tertiary and off-line storage.

  A diagram of magnetic disk is given next,

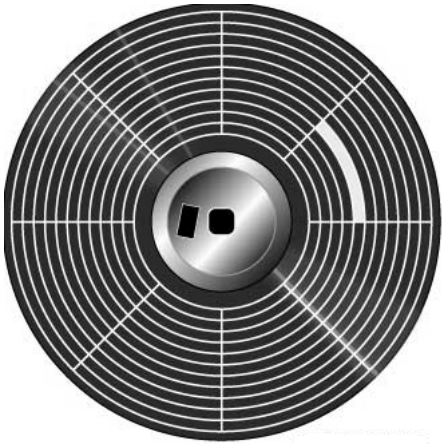


Figure 3. Magnetic disk.

- Semiconductor storage: Semiconductor memory uses semiconductor-based integrated circuits to store information. A semiconductor memory chip may contain millions of tiny transistors or capacitors. Both volatile and non-volatile

forms of semiconductor memory exist. In modern computers, primary storage almost exclusively consists of dynamic volatile semiconductor memory or dynamic random access memory. Since the turn of the century, a type of non-volatile semiconductor memory known as flash memory has steadily gained share as off-line storage for home computers. Non-volatile semiconductor memory is also used for secondary storage in various advanced electronic devices and specialized computers. The diagram given next shows the schematic of such drive,

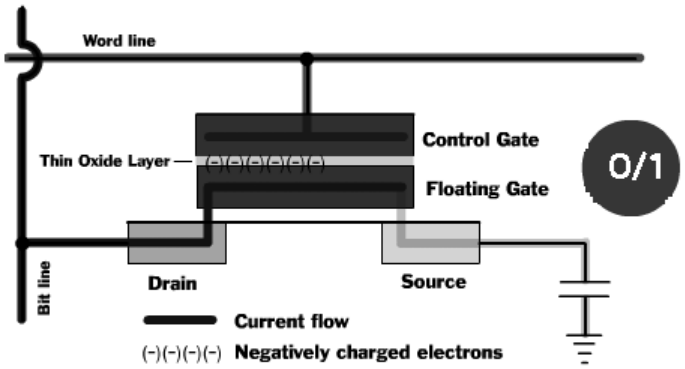


Figure 4. Schematic of semiconductor type storage.

- Optical disc storage: Optical disc storage uses tiny pits etched on the surface of a circular disc to store information, and reads this information by illuminating the surface with a laser diode and observing the reflection. Optical disc storage is non-volatile and sequential access. The following forms are currently in common use,

    - CD, CD-ROM, and DVD: Read only storage, used for mass distribution of digital information.
    - CD-R, DVD-R, DVD+R: Write once storage, used for tertiary and off-line storage.
    - CD-RW, DVD-RW, DVD+RW, DVD-RAM: Slow write, fast read storage, used for tertiary and off-line storage.
    - Blue-ray
    - HD DVD

A diagram of optical disk drive meant to read CD's is given next,

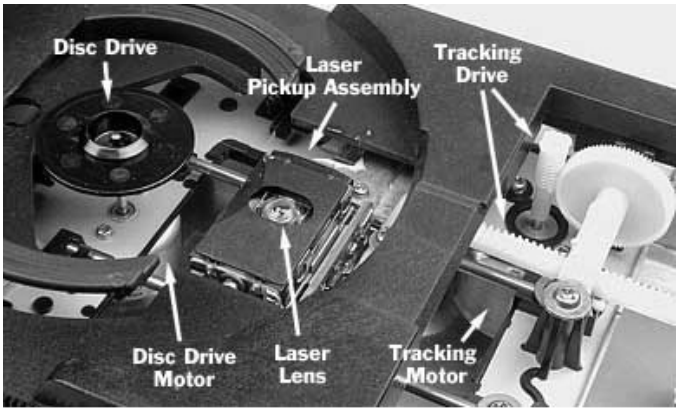


Figure 5. Optical Disc reading device.

- The following form have also been proposed:
  - HVD
  - Phase-change Dual

- Magneto-optical disc storage: Magneto-optical disc storage is optical disc storage where the magnetic state on a ferromagnetic surface stores information. The information is read optically and written by combining magnetic and optical methods. Magneto-optical disc storage is non-volatile, sequential access, slow write, fast read storage used for tertiary and off-line storage.

- Other early methods: Paper tape and punch cards have been used to store information for automatic processing since the 1890s, long before general-purpose computers existed. Information was recorded by punching holes into the paper or cardboard medium, and was read by electrically (or, later, optically) sensing whether a particular location on the medium was solid or contained a hole. A diagram of punch card is given next,

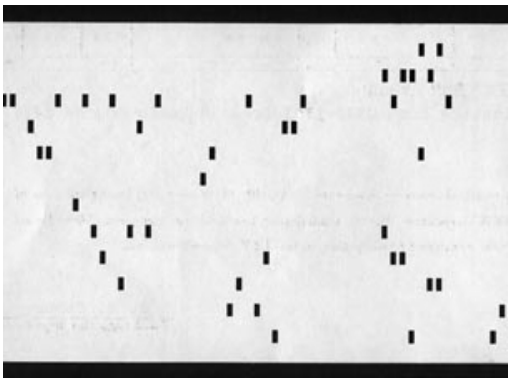


Figure 6. Punch card.

- Other proposed methods: -

    - Phase-change memory uses different mechanical phases of phase change material to store information, and reads the information by observing the varying electric resistance of the material. Phase-change memory would be non-volatile, random access read/write storage, and might be used for primary, secondary, and off-line storage.

    - Holographic storage stores information optically inside crystals or photopolymers. Holographic storage can utilize the whole volume of the storage medium, unlike optical disc storage, which is limited to a small number of surface layers. Holographic storage would be non-volatile, sequential access, and either write once or read/write storage. It might be used for secondary and off-line storage.

    - Molecular memory stores information in polymers that can store electric charge. Molecular memory might be especially suited for primary storage.

Some of the important aspects of any storage medium are presented below,

- Ability to change information.

- Addressability of information

    - In location-addressable storage, each individually accessible unit of information in storage is selected with its numerical memory address. In modern computers, location-addressable storage usually limits to primary storage, accessed internally by computer programs, since location-addressability is very efficient, but burdensome for humans.

    - In file system storage, information is divided into files of variable length, and a particular file is selected with human-readable directory and file names. The underlying device is still location-addressable, but the operating system of a computer provides the file system abstraction to make the operation more understandable. In modern computers, secondary, tertiary, and off-line storage use file systems.

    - In content-addressable storage, each individually accessible unit of information is selected with a hash value, or a short identifier with no pertaining to the memory address the information is stored on. Content-addressable storage can be implemented using software (computer program) or hardware (computer device), with hardware being faster but more expensive option.

- Capacity and performance.

- Storage capacity is the total amount of stored information that a storage device or medium can hold. It is expressed as a quantity of bits or bytes.

- Storage density refers to the compactness of stored information.

- Latency is the time it takes to access a particular location in storage. The relevant unit of measurement is typically nanosecond for primary storage, millisecond for secondary storage, and second for tertiary storage. It may make sense to separate read latency and write latency, and in case of sequential access storage, minimum, maximum and average latency.

- Throughput is the rate at which information can read from or written to the storage. In computer storage, throughput is usually expressed in terms of megabytes per second or MB/s, though bit rate may also be used. As with latency, read rate and write rate may need to be differentiated.

While making a decision about choosing the right storage medium one has to decide about the amount of the data the person wants to store. Various objects occupy various storage spaces and a list of amount of space a particular type of data may take, is given below,

- Bits
  - 1 bit: A binary decision.
- Bytes (8 Bits)
  - 1 byte: A single character.
  - 10 bytes: A single word.
  - 100 bytes: A sentence.
- Kilobyte (1024 Bytes)
  - 1 Kilobyte: A page of text.
  - 10 Kilobytes: A simple web page.
  - 100 Kilobytes: A compressed computer image or a long essay.
- Megabyte (1024 Kilobyte)
  - 1 Megabyte: A small novel.

  - 2 Megabytes: A high-resolution photograph.
  - 5 Megabytes: A very large volume of text.
  - 10 Megabytes: A minute of high-fidelity sound.
  - 100 Megabytes: More than 1 million pages of text.
  - 500 Megabytes: A CD-ROM.

- Gigabyte (1024 Megabyte)
  - 1 Gigabyte: A symphony in high-fidelity sound or a movie at TV quality.
  - 2 Gigabytes: More than 1 million pages of text.
  - 10 Gigabytes: A heavy music collection.
  - 20 Gigabytes: A VHS tape used for digital data.
  - 50 Gigabytes: A floor of books.
  - 100 Gigabytes: A floor of academic journals.

- Terabyte (1024 Gigabyte)
  - 1 Terabyte: 50000 trees made into paper and printed.
  - 2 Terabytes: An academic research library.
  - 10 Terabytes: The printed collection of the US Library of Congress.
  - 100 Terabytes: The entire Internet.

- Petabyte (1024 Terabyte)
  - 1 Petabyte: 3 years of EOS data.
  - 10 Petabytes: All US academic research libraries.
  - 100 Petabytes: All printed material.

- Exabyte (1024 Petabyte)
  - 1 Exabyte: All words ever spoken by human beings.

- Zettabyte (1024 Exabyte)

- Yottabyte (1024 Zettabyte)
- 1 Yottabyte: Everything that there is in the universe.

Various storage media come in various lifetime and capacity and depending amount the amount of the bytes needed to be stored the media is to be decided. For example, a normal CD can store upto 700MB data, DVD can store upto 8GB of data, HDD's can range between 4GB to 1TB etc- In recent years though the storage capacity has increased it has not been able to parallel itself with the growth rate of information. The diagram given below shows how the optical storage capacity has changed in recent years and now a day is serving as the most widely used medium of storage.

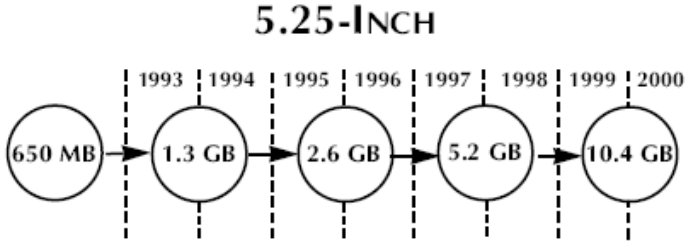


Figure 7. Timeline of storage capacity of 5.25" Optical Device.

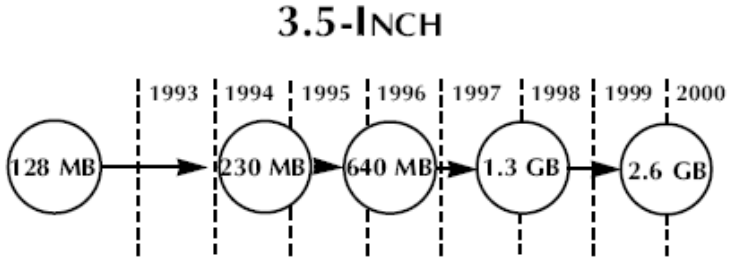


Figure 8. Timeline of storage capacity of 3.5" Optical Device.

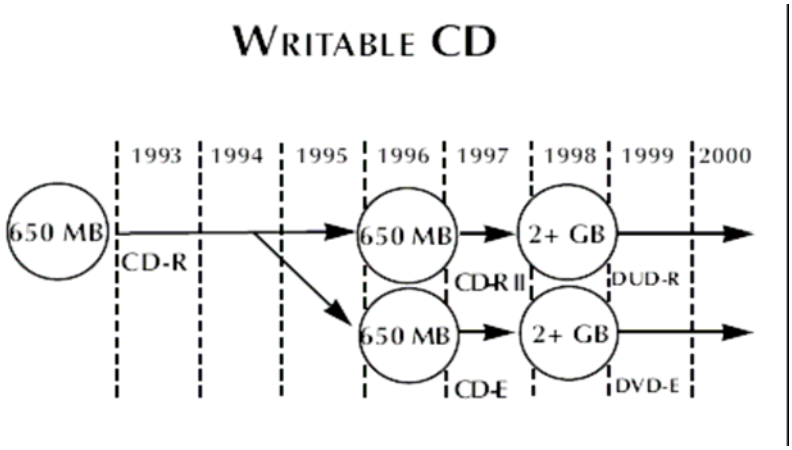


Figure 9. Timeline of storage capacity of Writable CD.

The storage capacity of various mediums is been standardized to spread uniformity. The standard capacity of the optical drives is given next,

- ISO 13963 - 230MB 3.5 inch Rewritable
- DIS 15041 - 640MB 3.5 inch Rewritable
- ISO 13549 - 1.3GB 5.25 inch Rewritable and WORM
- DIS 14517 - 2.6GB 5.25 inch Rewritable and WORM

Since, storage technologies mean different in context for different users, it can be divided into two parts i.e. enterprise and individual user. Based on the type of user the need varies and is to be identified before deciding upon the media of storage.

The table given next briefs the advantages, disadvantages and application of various storage media for individual purpose,

Technology	Advantages	Limitations	Applications
Compact disc, recordable (CD-R) or rewritable (CD-RW) and DVD	Low cost per megabyte Unlimited capacity with multiple discs Portable Widely-supported I/O interfaces Can be formatted for different data formats Long life Immune to corruption once data is written (CD-R and DVD only)	Limited capacity on one disc (though much greater than diskette) Slow to moderate read/write speed	Data archiving Data distribution Data migration Localized file sharing Offsite storage
Diskettes, 1.44 MB	Simple to use	Limited	Local data transfer of

	Portable Can be formatted for different data formats	capacity Limited read/write speed Not supported by many newer computers	small files Storage of small files or programs
Hard drive, external	• High read/write speed • Can be moved among computers	• Limited capacity • Awkward for data transfer among multiple computers	• Local backup • Local archiving
Hard drive, internal	• Convenient; usually comes with the computer • High read/write speed • Convenient for use with single computer (but can be shared among multiple computers with proper support • Most common form of data storage	• Limited capacity • Without special support, confined to a single computer or server	• Storage in a single computer • Swap files
Removable storage (ZIP disks, JAZ disks, etc.)	• Simplicity • Portability • Unlimited capacity	• Proprietary media • Limited read/write	• Personal computing • Local data transfer of

	with multiple disks • Convenient for use with single computer	speed • High cost per megabyte	small files • Local backup • Local archiving
Solid-state storage (USB devices, flash memory, smart cards, etc.)	• No mechanical parts • High read/write speed • Small form factor	• Limited storage capacity • High cost per I/O operation	• Swap files • Local data transfer • Internet service providers • Video processing • Relational databases • High-speed data acquisition

Table 5. Advantages, disadvantages and application of various storage mediums for an individual.

The table given next briefs the advantages, disadvantages and application of various storage media for enterprise purpose,

Technology	Advantages	Limitations	Applications
Direct-attached storage (DAS)	• Simplicity • Low initial cost • Ease of management	• Storage for each server must be administered separately • Inconvenient for data transfer in network environments • Server bears load of processing applications	• Data and application sharing • Data backup • Data archiving